Seattle Bird Call Classification Using Neural Networks

1. ABSTRACT

This report explores how deep learning can be used to classify bird species based on their sounds. It uses mel spectrogram data from 12 Seattle-area bird species and applies Convolutional Neural Networks (CNNs) to identify patterns in the calls. The project tests both binary and multi-class classification models and compares their performance. Results show that simpler (Lighter) CNN architectures performed better, especially when dealing with unbalanced training data. Final model predictions on new audio clips confirm the model's ability to generalize to unseen sounds.

1.2 INTRODUCTION

This project uses audio data from 12 common bird species found in the Seattle area. Each audio clip contains bird calls, which were converted into **Mel Spectrograms**; a visual form of sound that makes it easier to recognize patterns using machine learning. These spectrograms were provided in a pre-processed format for model training.

The main goal of this project is to classify bird species using these spectrograms. Both **Binary classification** (used to distinguish between two species) and **Multi-Class classification** (used to predict one of twelve possible species) were explored. Various Convolutional Neural Network (CNN) architectures were tested and compared. The final part of the project uses the best-performing model to predict species from three new, unlabeled bird sound recordings.

2. THEORETICAL BACKGROUND

This project uses Convolutional Neural Networks (CNNs), a type of deep learning model that works especially well with images. Since bird calls can be shown as mel spectrograms, CNNs are a good fit for this task.

A mel spectrogram shows how sound frequencies change over time, and it adjusts the scale to match how humans (and animals) hear. By using these spectrograms, we allow the CNN model to learn patterns that are unique to each species of bird.

The CNN models in this project use multiple layers, including convolutional layers that detect small features in the spectrograms, pooling layers that reduce size and keep important information, and dropout layers that prevent overfitting. For the multi-class task, we also used class weighting to help balance the training process since some species had fewer samples.

3. METHOLOGY

The dataset used in this project includes mel spectrograms of bird calls from 12 different bird species found in the Seattle area. Each sample is a 2 second segment converted into a spectrogram with a fixed shape of 128 x 517, representing frequency and time.

Before training any model, the dataset was processed as follows:

- Each spectrogram was normalized to ensure consistent scaling across inputs.
- Data was reshaped into the format required by CNNs: (samples, 128, 517, 1).
- The data was shuffled and split into training (70%) and testing (30%) sets.
- For the multi-class task, class weights were calculated to handle class imbalance.

This report explores three types of CNN architecture:

- Base CNN (v1): A small network with two convolutional layers and dropout.
- Deeper CNN (v2): A more complex network with additional convolutional layers and regularization.
- Lighter CNN (v3): A smaller and faster model designed to avoid overfitting.

Each architecture was tested on two tasks:

- Binary classification: Classifying between two randomly selected species.
- Multi-class classification: Predicting which of the 12 species produced a given spectrogram.

All models were trained using the Adam optimizer with a learning rate of 0.001 and binary or categorical cross-entropy as the loss function. The number of epochs was chosen based on training time and convergence behavior.

Model performance for each part of the model was evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The best-performing model (lighter CNN) was then used to classify three formatted test audio clips.

4. RESULTS

Combined Classification Report (%)

	Model	Accuracy	Precision	Recall	F1-Score
0	Base CNN (v1)	57.75%	53.09%	57.75%	53.25%
1	Deeper CNN (v2)	61.97%	57.86%	61.97%	51.80%
2	Lighter CNN (v3)	70.42%	75.57%	70.42%	65.02%

FIG 1.: Binary Classification Report Table

The Lighter CNN (v3) outperformed both the Base and Deeper versions in all evaluation metrics.

The Base and Deeper CNNs struggled due to class imbalance. The Lighter CNN showed improved performance and stability.

Combined Model Performance Summary (%)

	Model	Accuracy	Precision	Recall	F1-Score
0	Multi-class Base CNN	2.35%	0.06%	2.35%	0.12%
1	Multi-class Deeper CNN	4.54%	2.12%	4.54%	0.92%
2	Multi-class Lighter CNN	26.89%	35.09%	26.89%	28.30%

FIG 2.: Multi-Class Classification Report Table

Predicted bird species for each test audio clip

	Test Clip	Top Prediction	Confidence (%)	Top 3 Predictions
0	Test1	sonspa	11.87%	sonspa, rewbla, houfin
1	Test2	rewbla	31.79%	rewbla, whcspa, houspa
2	Test3	sonspa	31.77%	sonspa, rewbla, houfin

FIG 3.: Final Prediction Report

The model was able to make consistent predictions, with **'sonspa'** and **'rewbla'** appearing across all 3 test cases.

This confusion matrix shows the performance of the Lighter CNN on the binary classification task. Out of 44 actual Class 0 samples, the model correctly predicted 43 and misclassified 1. For Class 1, it correctly identified 7 samples but misclassified 20 as Class 0.
While the model was very accurate with Class 0, it struggled more with Class 1, possibly due to class imbalance or overlapping features between the two species.

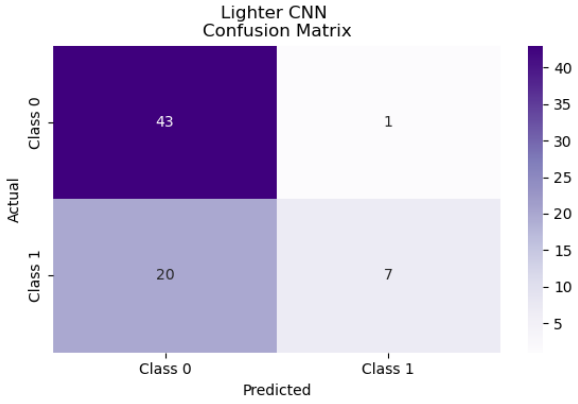

FIG 4.: Confusion Matrix Binary Lighter CNN

5. DISCUSSION

This project tested different CNN models to classify bird species using mel spectrograms. The Lighter CNN gave the best results in both binary and multi-class classification. It also trained faster and was more stable than the deeper models.

There were some limitations that affected the performance. The dataset was unbalanced, some bird species had too many samples, while others had very few. Because of this, the model was better at recognizing the species with more data and often misclassified the ones with less. Some species also had very similar-sounding calls. When listening to the clips and looking at the spectrograms, their patterns looked alike, which likely confused the model.

Another issue was the lack of data variety. No data augmentation was used, which might have helped the model learn better. Also, the test clips included background noise or more than one bird, making it harder to be sure about the predictions.

Even with these challenges, CNNs were a good choice for this task because spectrograms are like images. CNNs can find patterns in these images without needing us to manually pick features. Overall, the project showed that CNNs can work well for bird sound classification, even with some limitations in the dataset.

6. LIMITATIONS

Class Imbalance:

Some bird species had many more samples than others, which made it difficult for the model to learn equally across all classes. The underrepresented species often had poor recall and were more likely to be misclassified.

Similar Sound Patterns:

Many species have overlapping frequency ranges and similar vocal patterns, which made it hard for the model to distinguish between them, especially when calls were noisy or unclear in the spectrogram.

Limited Data Per Class:

Although spectrograms were generated, the overall number of usable samples per species was still relatively small. This limited the model's ability to generalize and may have contributed to overfitting in deeper CNNs.

Lack of Data Augmentation:

No data augmentation (like pitch shift, noise addition, or time stretch) was used. This could have helped the models generalize better by exposing them to a wider variety of input variations.

7. FUTURE WORK

- **Adding Data Augmentation:**
  Techniques like noise addition, time stretching, or pitch shifting can improve model learning and make it more robust to different types of input.

- **Advanced or pre-trained models:**
  Using architectures like transfer learning with pretrained CNNs could help improve accuracy, especially for multi-class task.

- **Balancing the dataset:**
  Collecting more samples for underrepresented species would help the model learn each class more evenly and improve prediction performance.

- **Testing in real-world settings:**
  Running the model on longer or live audio clips could help evaluate how well it performs outside of the trimmed dataset.

8. CONCLUSION

This project explored how Convolutional Neural Networks (CNNs) can be used to classify bird species based on audio recordings. By converting bird calls into mel spectrograms, we were able to apply image-based deep learning techniques to an audio classification task.

Out of the three CNN models tested, the Lighter CNN performed the best in both binary and multi-class tasks, offering better accuracy and faster training. The project also showed how well the model could generalize to new, unseen audio clips, even with some limitations in the data.

Overall, this work showed that CNNs are a good option for bird call classification, especially when combined with proper data preparation and thoughtful model design.

9. PROBLEMS FACED AND SOLUTIONS

Throughout this project, I encountered several technical and practical challenges.

9.1 Class Imbalance

Some bird species had very few samples compared to others. This made it difficult for the model to learn all classes equally.

Solution: Used class weights during training to reduce the impact of imbalance in the multi-class model.

9.2 Model Overfitting

The deeper CNN model started overfitting quickly, especially during multi-class model training.

Solution: Used dropout layers and also added ***'early stopping'*** (a callback function) to stop training when the model's performance on validation data stopped improving.

9.3 Similar Sound Patterns

Certain bird calls looked very similar in spectrograms, causing the model to confuse them.

Solution: Included confusion matrices and classification reports to better analyze where and why these errors were happening.

9.4 Difficulty Interpreting Output

The test clips had background noise and possibly more than one bird, leading to lower prediction confidence.

Solution: Limited the clip duration and used normalized spectrograms to improve input quality.

10. REFERENCES

- https://www.xeno-canto.org/
- https://librosa.org/doc/latest/index.html
- https://keras.io
- https://scikit-learn.org
- https://docs.python.org/3/
- https://www.youtube.com/watch?v=woEs7UCaITo
- https://www.youtube.com/watch?v=2Fp1N6dof0Y
- https://www.youtube.com/watch?v=YRhxdVk_sIs

11. GITHUB LINK: https://github.com/mehek1708/STML2_HW3.git